Unsupervised Model Selection for Recognition of Regional Accented Speech

Najafian, Maryam; DeMarco, Andrea; Cox, Stephen; Russell, Martin

Identification of Age-Group from Children's Speech by Computers and Humans

Saeid Safavi, Martin Russell and Peter Jančovič

School of Electronic, Electrical & Computer Engineering, University of Birmingham, UK

{sxs796, m.j.russell, p.jancovic}@bham.ac.uk

Abstract

This paper presents results on age-group identification (Age-ID) for children's speech, using the OGI Kids corpus and GMM-UBM, GMM-SVM and i-vector systems. Regions of the spectrum containing important age information for children are identified by conducting Age-ID experiments over 21 frequency sub-bands. Results show that the frequencies above 5.5 kHz are least useful for Age-ID. The effect of using gender-independent and gender-dependent age-group modelling is explored. The GMM-UBM and i-vector systems considerably outperform the GMM-SVM system. The best Age-ID performance of 85.77% is obtained by the i-vector system applied to band-limited speech to 5.5 kHz. Experiments on human Age-ID were also conducted and the results show that the humans do not achieve the performance of the machine.

Index Terms: paralinguistic speech processing, age identification, child speech, gaussian mixture model, support vector machine, i-vector, frequency band

1. Introduction

In addition to its linguistic content, the acoustic speech signal also contains paralinguistic information, such as the speaker's identity, accent, gender, age, or emotional state. Automatic recognition of such paralinguistic information for children can be beneficial in many application areas. It could be employed to adapt speech models, to guide a child computer interaction system to automatically adapt content, to enhance child security and protection or in wide range of educational applications. For instance, some social networking sites are designed specifically for children, (for example, "Club Penguin"[1]). As such systems evolve to include speech, an automatic system that recognises the age, gender, or identity of a person from his or her voice could be a valuable safeguard for a child engaged in social networking, for example to provide protection from an adult masquerading as a child. In education, an interactive educational tutor could recognise the age of a child and adapt its content appropriately.

Research in paralinguistic speech processing has grown considerably over the last two decades. It was initially focused on adults speech (for example, [1]), but more recently it has expanded to include children's speech [2] [3]. The task of age and gender recognition from adults' speech is a particular area of paralinguistic speech technology that has received attention [4, 5, 6, 7]. Research focused on exploring the use of features capturing different types of information from the speech signal and the use of different machine learning algorithm. Most studies employed mel-frequency cepstral coefficients (MFCCs). The use of TempoRAl PatternS (TRAPS) as features to capture longer temporal context was explored in [6]. Several studies also considered the use of glottal and prosodic features. These were typically calculated on the whole utterance and included features such as the fundamental frequency, jitter/shimmer, articulation rate, and harmonic-to-noise ratio [4, 5, 6].

Previous research [8] has demonstrated that the information that is most useful for, say, speaker-identification lies in different frequency bands to that which is most useful for accent-identification. Similarly, it seems likely that the information that is useful for age detection will not be uniformly distributed across the spectrum.

Recent advances based on the principle of factor analysis (for example [9]), have improved classification accuracy. In this method i-vectors, which are a compact representation of an utterance in the form of a low-dimensional feature vector, are used as a new feature set instead of high dimensional supervectors. These i-vectors determine the principal components of the total variability space. In [10] the use of i-vectors as features, and a SVM classifier as the decision maker, was studied for age recognition from adults' speech. The same idea was also applied by [11] for spoken language recognition. During this research the effects of using different machine learning algorithms and scoring techniques were also investigated.

The Age-ID task is to predict the age of a speaker from a sample of speech from that speaker. This can be carried out in a classification scenario [12] using age groups, or by using regression [13], i.e., predicting the age in years. As the OGI Kids speech corpus only provides the information about the school grade of each child and not the actual age, we focus on age group identification from children's speech. In this paper the results of experiments on Age-ID from children's speech are presented. In addition to investigating the use of conventional GMM-UBM and GMM-SVM systems, we replaced GMM mean-supervectors by low-dimensional i-vectors to model utterances in a total variability space and we use these vectors as a new feature sets for identifying the age-group of each test speaker.

2. The OGI kids' speech corpus and data description

The OGI Kids Speech corpus [14] is a collection of spontaneous and read speech recorded at the Northwest Regional School District near Portland, Oregon. As described in [14] the toolkit form Center for Spoken Language Understanding (CSLU) is used for data collection. A gender-balanced group of approximately 100 children per grade from Kindergarten (5-6 year olds) through to grade 10 (15-16 year olds) participated in the collection. For each utterance, the text of the prompt was displayed on a screen, and a recording of a person speaking the prompt

[1] http://www.clubpenguin.com/company/about

Table 1: The Center Frequencies for 24 Mel-spaced Band-Pass Filters

FILTER NUMBER	CENTER FREQ. (Hz)	FILTER NUMBER	CENTER FREQ. (Hz)
1	156	13	1843
2	281	14	2062
3	406	15	2343
4	500	16	2656
5	625	17	3000
6	750	18	3375
7	875	19	3812
8	1000	20	4312
9	1125	21	4906
10	1281	22	5531
11	1437	23	6281
12	1625	24	7093

was played, in synchrony with facial animation using the animated 3D character "Baldi". The subject then repeated the prompt, which was recorded via a head-mounted microphone and digitized at 16 bits and 16 kHz. In total the corpus comprises recordings of words and sentences from 1100 speakers. In this study, 766 speakers were chosen randomly for testing and the remaining 334 for training. For age-group identification, the age groups are specified as:

AG1: kindergarten to 3^{rd} grade (5-9 year olds),

AG2: 4^{th} to 7^{th} grade (9-13 year olds), and

AG3: 8^{th} to 10^{th} grade (13-16 year olds).

The individual age groups AG1, AG2, and AG3 contained 290, 285 and 191 test speakers, respectively.

3. Age identification systems

3.1. Signal analysis

Feature extraction was performed as follows. Periods of silence were discarded using an energy-based Speech Activity Detector. The speech was then segmented into 20 ms frames (10 ms overlap) and a Hamming window was applied. The short-time magnitude spectrum, obtained by applying an FFT, is passed to a bank of 24 Mel-spaced triangular bandpass filters, spanning the frequency region from 0 Hz to 8000 Hz. Table 1 shows the center frequency of each filter (the cut-off frequencies of a filter are the centre frequencies of the adjacent filters). To investigate the effect of different frequency regions on Age-ID performance, experiments were conducted using frequency band limited speech data comprising the outputs of groups of 4 adjacent filters. We considered 21 overlapping sub-bands, where the N^{th} sub-band comprises the outputs of filters N to $N+3$ (N=1 to 21). Each set of 4 filter outputs was transformed to 4 Mel Frequency cepstral coefficients (MFCCs) plus 4 delta and 4 delta-delta MFCCs, and feature warping [15], using short-time gaussianization was applied. For the full bandwidth experiments the outputs of all 24 filters were transformed into 19 static plus 19 delta and 19 delta-delta MFCCs.

3.2. Modelling

Our Age-ID systems are based on the GMM-UBM [16, 17], GMM-SVM [17] and factor analysis (using i-vectors as features) [11] methods.

In the GMM-UBM approach, a UBM is built using all utterances from 334 speakers. The gender-independent age group models are obtained by MAP adaptation (adapting the means only) of the UBM, using the age-group-specific training data. The result is one UBM and 3 age-group GMMs. To investigate the effect of using gender-dependent age-group models on Age-ID performance, gender-dependent models are obtained by MAP adaptation of the UBM, using the gender and age-group-specific training data.

In our GMM-SVM system, the mean vectors of MAP-adapted GMMs of each age-group (obtained as described above) were concatenated into a supervector [17]. The age-group classes are assumed to be linearly separable in the supervector space. The supervectors are used to build one SVM for each age-group, by treating that age-group as the target class and the others as the background classes.

In the related field of speaker recognition, the combination of generative (GMM-UBM) and discriminative (GMM-SVM) modelling approaches, based on GMM means supervectors, have been shown to provide a good level of performance [16, 17]. However recent progress has found an alternate method of modeling GMM supervectors that provides superior speaker recognition performance [9]. This technique is referred to as total variability modeling. Total variability modeling assumes the GMM mean supervector, μ, that best represents a set of feature vectors can be decomposed as

$$\mu = m + Tw \tag{1}$$

where m is the mean supervector of the UBM, T spans a low-dimensional total variability subspace and w is a vector that best describe the utterance dependent mean offset Tw. The vector w is commonly referred to as the i-vector and has a standard normal distribution $N(0, I)$ and T is the rectangular low rank matrix which is estimated via factor analysis. In the total variability modeling approach, i-vectors are the low-dimensional representation of an audio recording that can be used for classification and estimation purposes.

The scoring approach proposed in [11] for language identification, is used in this research. Linear discriminant analysis (LDA) is used to find a new basis for the total variability space such that for any D, the subspace spanned by the first D LDA basis vectors maximises the between-class variability while minimising the within-class variability. LDA is applied to the i-vectors for all training data from all age-groups and defines a projection matrix A^t from the total variability space onto the subspace spanned by the first D LDA basis vectors. D is usually set to $Q-1$, where Q is the number of classes.

In this system each age-group c is then represented by the mean of projected and length normalized i-vectors for that class as

$$m_c = \frac{\sum_{j=1}^{N_c} \tilde{w}_j}{||\sum_{j=1}^{N_c} \tilde{w}_j||} \tag{2}$$

where N_c is the total number of utterances for the age-group c and the unit normalized LDA i-vectors are

$$\tilde{w}_j = \frac{A^t w_j}{||A^t w_j||} \tag{3}$$

where w_j is the extracted i-vector from a utterance j.

At the recognition stage, the score for each class c is calculated as the dot product of the unit normalized LDA test i-vector with the age-group model mean, i.e.,

$$score(c) = \tilde{w}_{test}^T m_c. \tag{4}$$

This score expresses the angle between the unit normalized LDA test i-vector and the mean of projected and length normalized i-vectors for each class. The overall block diagram of the i-vector based identification system is illustrated in Figure 1.

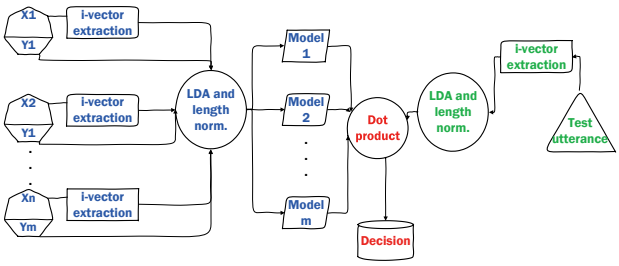

Figure 1: *The block diagram of the Age-ID system based on the i-vector approach, depicting both training and testing phase. X_n and Y_n represent samples and class labels, respectively.*

The use of i-vectors for age estimation has several distinct advantages over GMM supervectors, for example the relatively low dimensionality of i-vectors significantly reduces the computational cost of model training and estimation compared to a GMM supervector system. Thus the method lends itself to real-time implementation, which is important for applications.

4. Experimental results and discussion

4.1. Age-ID for children's speech using isolated frequency sub-bands

In this section, we study the effect of different sub-bands on Age-ID performance for children's speech. Experiments are conducted separately on 21 sub-bands, each consisting of four consecutive channels (see Section 3.1 for more details), and using the gender independent GMM-UBM system. For each sub-band, three gender independent age-group models are trained, corresponding to AG1, AG2 and AG3. The models have 64 mixture components, which was found to be adequate for these 12 dimensional sub-band features.

Figure 2 presents the average Age-ID results as a function of frequency sub-band. It is evident that the performance even when using a narrow frequency region is in most cases well above chance. Figure 3 contrasts the usefulness of sub-bands for Age-ID and Gender-ID. The figure was obtained by normalising the data in Figure 2 so that the sum of the values over all of the sub-bands is 1. The same procedure was then applied to the corresponding sub-band results on Gender-ID presented in [3]. The normalised Age-ID results were then subtracted from the normalised Gender-ID results to obtain Figure 3 (similar procedure is described in [8]). Thus, negative regions in Figure 3 indicate sub-bands which are more useful for Age-ID, while positive values indicate sub-bands that are useful for Gender-ID. The results indicate that the most useful sub-bands for Age-ID, in comparison to Gender-ID, are the sub-bands 3 and 4 (281 Hz to 625 Hz), and from 13 to 16 (1.62 kHz to 3 kHz). Thus, while Gender-ID appears to make use of similar information to speaker recognition, Age-ID is more similar in these respects to speech recognition or accent ID [8].

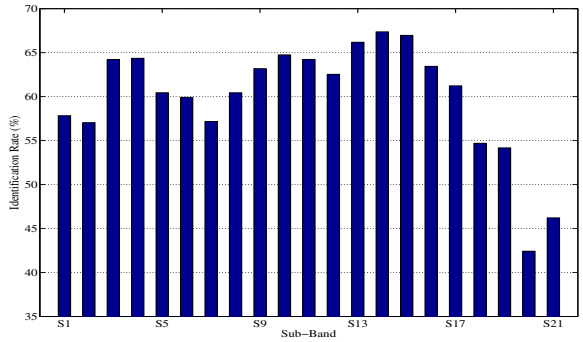

Figure 2: *The effect of different frequency sub-bands on Age-ID, average identification rate over all age groups.*

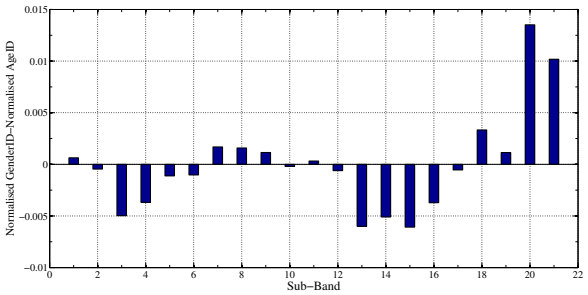

Figure 3: *The difference between the normalized Gender-ID and Age-ID performance for frequency sub-bands.*

4.2. Age-ID using full/restricted bandwidth speech

This section presents the Age-ID results that are obtained using the GMM-UBM, GMM-SVM and i-vector based systems described in section 3.2. Experiments were performed using full-bandwidth (FB) and band-limited speech (BL). The band-limited case includes frequencies up to 5.5 kHz, which corresponds to the frequency region covered by all sub-bands except sub-bands 18 to 21.

We first study the effect of using gender dependent and independent age-group modeling, using the GMM-UBM system. The results of this study are shown in the first two rows of Table 2. For Age-ID, gender-independent modeling gives better results than gender-dependent modeling. This could be due to the amount of training data, as in the case of gender dependent age modeling we split all training data into 6 groups, compare to gender independent age modeling which training data split into three groups.Based on these results, subsequent experiments use gender-independent modelling.

Then, we demonstrate the effects of employing the discriminative GMM-SVM and i-vector systems when using gender-independent modelling. For each of the systems, we performed experiments using different numbers of mixture components. The best results were obtained when using 1024, 512 and 256 mixture components for the GMM-UBM, GMM-SVM and i-vector system, respectively, and these are presented in Table 2. For the i-vector system, we performed experiments using different numbers of dimensions for training the total variability matrix. The best results were obtained using 400 dimensions for the T matrix. It can be seen that the i-vector system out-

performs the GMM-UBM and GMM-SVM systems, especially when band limited speech is used. Table 3 shows a confusion matrix obtained by the i-vector system using band-limited speech. Each row corresponds to a grade and shows the percentages of children in that grade that were classified as being in AG1, AG2 and AG3. The dotted lines indicate the boundaries of AG1, AG2 and AG3. The top and bottom halves of the table correspond to male and female speakers, respectively. The table shows similar characteristics for boys and girls up to 7^{th} grade, with the majority of errors near age-group boundaries. At the boundary between AG1 and AG2, 10% of 3^{rd} grade boys (AG1) are incorrectly classified as AG2 and 33% of 4^{th} grade boys (AG2) are incorrectly classified as AG1. For girls the corresponding figures are 8% and 39%. For 7^{th} grade (AG2) 45% of boys and 24% of girls are classified as being in AG3, while for 8^{th} grade (AG3) 29% of female speakers are classified as AG2 but none of the boys are misclassified. The inconsistency between the results for boys and girls at the AG2-AG3 boundary may be because AG3 contains speech from a number of boys whose voices have broken. It may be that gender-dependent modelling is needed for AG3, even though it is not advantageous overall, or that, as in the case of gender identification [3], it is necessary to build separate models for AG3 boys whose voices have and have not broken.

Table 2: *Age-ID recognition rate (in %) obtained by the gender-independent GMM-UBM, GMM-SVM and i-vector systems and gender-dependent GMM-UBM system.*

System	Age-ID rate (%)	
	Full-bandwidth	Band-limited
GMM-UBM (gender dep.)	71.76	-
GMM-UBM (gender indep.)	82.01	84.07
GMM-SVM (gender indep.)	79.77	-
i-vector (gender indep.)	**82.62**	**85.77**

Table 3: *Confusion matrix for age identification (in %) for three age groups, obtained by the i-vector system using band-limited speech.*

Grade-index	Model-index AG1 (%)	AG2 (%)	AG3 (%)
Male			
k	100	0	0
1^{st}	100	0	0
2^{nd}	97.43	2.56	0
3^{rd}	85.71	10.20	4.08
4^{th}	33.33	60.60	6.06
5^{th}	8.57	82.85	8.57
6^{th}	6.97	81.39	11.62
7^{th}	0	54.83	45.16
8^{th}	0	0	100
9^{th}	2.17	6.52	91.30
10^{th}	0	0	100
Female			
k	100	0	0
1^{st}	100	0	0
2^{nd}	97.87	2.12	0
3^{rd}	92.10	7.89	0
4^{th}	38.70	61.29	0
5^{th}	11.42	82.85	5.71
6^{th}	10.00	80.00	10.00
7^{th}	2.70	72.97	24.32
8^{th}	0	29.03	70.96
9^{th}	5.00	20.00	75.00
10^{th}	0	30.00	70.00

4.3. Human Age-ID for children's speech

In addition to the computer Age-ID experiments presented in the previous sections, we also performed experiments on Age-ID by human listeners. The test set consisted of the same 766 test utterances used in the computer Age-ID experiments. Twenty listeners participated in the experimental evaluations. Each participant listened to 38 utterances on average. The length of each utterance was 10 seconds. All human listening tests were performed in a quiet room using the same PC and a high quality headphones. The Age-ID rates for each age-group achieved by human listeners are presented in Table 4. The average performance over all age groups was 67.54%.

Table 4: *Confusion matrix for age identification (in %) for three age groups, obtained by human listeners.*

Test-index	Model-index AG1	AG2	AG3
AG1	**81.2**	16.9	1.8
AG2	25.5	**50.8**	23.6
AG3	3.8	24.4	**71.7**

5. Conclusions

To conclude, our results for Age-ID based on narrow frequency sub-bands indicate that the performance, even for narrow frequency regions, is in most cases well above chance. Moreover, a comparison of useful bands for Age-ID and Gender-ID shows that most of the useful information for Age-ID is in similar regions of the spectrum to those that are useful for speaker ID. This result suggests that removing higher parts of the spectrum will improve Age-ID performance. Hence we compared Age-ID performance for full-bandwidth (up to 8 kHz) and restricted bandwidth (up to 5530 Hz). As expected, performance is improved for both the GMM-UBM and i-vector systems when band-limited speech is used. The best Age-ID performance is 87.55% obtained from the i-vector system applied to band-limited speech.

Further analysis of the results from the best system shows that Age-ID for young children, for both male and female speakers, is a relatively easy task. The main confusion arises with male and female speakers who belong to the 4^{th} and 7^{th} grades. These grades are at the boundary of AG2. For example, 33.33% and 38.70% of 4^{th} grade boys and girls are miss identified as belonging to AG1, respectively. It is also evident that for AG3, Age-ID for girls is more challenging than for boys from same age-group.

6. Acknowledgements

Thanks to Microsoft conversational systems research center, specifically Seyed Omid Sajadi, [18] for providing the MSR toolbox. During this research some functions from this toolbox is modified and used along with other self written codes.

7. References

[1] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language - State-of-the-art and the challenge," *Computer Speech Language*, vol. 27, no. 1, pp. 4–39, 2013.

[2] S. Safavi, M. Najafian, A. Hanani, M. Russell, P. Jančovič, and M. Carey, "Speaker recognition for children's speech," *Interspeech*, 2012.

[3] S. Safavi, P. Jančovič, M. J. Russell, and M. J. Carey, "Identification of gender from children's speech by computers and humans." in *INTERSPEECH*. ISCA, 2013, pp. 2440–2444.

[4] C. Müller, "Automatic recognition of speakers' age and gender on the basis of empirical studies," *Interspeech*, pp. 2118–2121, 2006.

[5] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Müller, R. Huber, B. Andrassy, J. G. Bauer, and B. Littel, "Comparison of four approaches to age and gender recognition for telephone applications," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[6] T. Bocklet, G. Stemmer, V. Zeissler, and E. Noth, "Age and gender recognition based on multiple systems - early vs. late fusion," *Interspeech*, pp. 2830–2833, 2010.

[7] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech Language*, vol. 27, no. 1, pp. 151–167, 2013.

[8] S. Safavi, A. Hanani, M. Russell, P. Jančovič, and M. Carey, "Contrasting the effects of different frequency bands on speaker and accent identification," *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 829–832, 2012.

[9] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[10] M. H. Bahari, M. McLaren, H. V. Hamme, and D. A. van Leeuwen, "Age estimation from telephone speech using i-vectors." in *INTERSPEECH*. ISCA, 2012.

[11] P. A. Torres-Carrasquillo, E. Singer, W. M. Campbell, T. P. Gleason, A. McCree, D. A. Reynolds, F. Richardson, W. Shen, and D. E. Sturim, "The mitll nist lre 2007 language recognition system." in *INTERSPEECH*. ISCA, 2008, pp. 719–722.

[12] T. Bocklet, A. Maier, and E. Nth, "Age determination of children in preschool and primary school age with gmm-based supervectors and support vector machines regression." in *TSD*, ser. Lecture Notes in Computer Science, vol. 5246. Springer, 2008, pp. 253–260.

[13] D. A. van Leeuwen and M. H. Bahari, "Calibration of probabilistic age recognition." in *INTERSPEECH*. ISCA, 2012.

[14] K. Shobaki, J. P. Hosom, and R. A. Cole, "The ogi kids' speech corpus and recognizers," *Int. Conf. on Spoken Language Processing*, 2000.

[15] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," *Speaker Odyssey*, 2001.

[16] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[17] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

[18] M. S. Seyed Omid Sadjadi and L. Heck, "MSR identity toolbox v1.0: A MATLAB toolbox for speaker-recognition research," *Speech and Language Processing Technical Committee Newsletter, IEEE*, 2013.